MULTIVARIATE DATA ANLYSIS PROJECT

Gabriel Bota Ochogavia, Arnau Pons Olivan, Adrian Quirante González

UPC DATA ANLYSIS

Index

Motivation

Data science requires techniques to extract insights from complex datasets. For our final project, we chose a Spanish wine dataset (dataset) to apply a variety of methods studied in class: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Correspondence Analysis (CA), Clustering, Quadratic Discriminant Analysis (QDA), and Linear Discriminant Analysis (LDA).

It is quite difficult to get one dataset which can be used for all these indicated techniques. Majority of datasets are not diverse enough to allow for a detailed analysis, a large portion of them consist of either numeric or categorical data. The Spanish wine dataset provides us many useful attributes as compared to the other datasets so that we can use this data for dimensionality reduction by using PCA, for visualizing the similarities between the samples by using MDS, for analyzing the categorical data by using CA and finally for classification by using QDA/LDA. It also naturally supports clustering because it often follows inherent grouping of wines by type and/or origin.

This is done not only for the practical application of these methods in the project but also to show that data analysis methods have significant uses when applied in a practical scenario.

Additionally, we aim to improve our communication skills by effectively presenting the results to diverse audiences with varying levels of data science knowledge, translating complex analytical findings into accessible insights. However, all the results obtained may be more elaborated in the RMD attached file (Appendix A).

The dataset

The selected data is suitable to work through all the methods addressed above given that it contains both numerical and categorical data type. In particular, the given table consists of five categorical variables like winery or region, and 6 numerical variables containing information on wine age, properties, and individual ratings. However, based on the technique utilized some of these variables can be taken as both numerical and categorical when fitting the model. There is a total of 7500 observations.

The vast majority of the methods, such as PCA, requires the input data to possess specific characteristics. For instance, PCA often works best when the numerical data conforms to a normal distribution or other distributions like binomial, exponentially, or gamma distributed data. Nonetheless, by looking at the distribution of the numerical variables we have chosen for our study, it is easy to notice that they do not meet these requirements. As in most other data science projects, it will be also necessary to preprocess our data before analysing it.

Preprocessing

In order to obtain better results from the methods, some preprocessing techniques will be applied to improve the properties of the input data. Among all these tools, we are going to focus on linear transforms of data and outlier treatment.

First of all, missing values are omitted. Then, variables where lower values cumulate much more probability than higher ones, are transformed with the logarithm to distribute the data. Then, in some variables, outliers are also removed.

For instance, let's consider the wine price as an example:

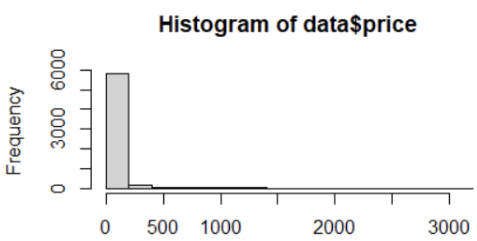

FIGURE 1

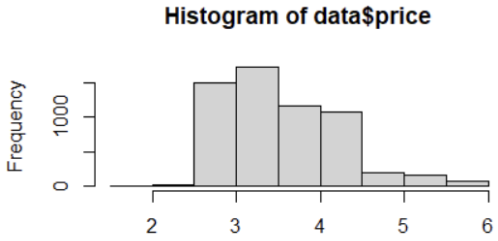

FIGURE 2

After preprocessing, although variables may not yet follow a gaussian distribution, applying techniques to the processed data yields better results.

PCA

PCA is a statistical technique that is used in data analysis in order to achieve a lower dimensional representation of the input data while retaining maximum variability of the data. This is achieved through a process of finding new variables named principle components which are orthogonal and are obtained as linear combinations of the original variables.

After applying PCA we obtain the following table:

Component	Eigenvalue	% variance	% cumulative variance
1	2.12	35.45	35.45
2	1.18	19.68	55.14
3	0.97	16.19	71.33
4	0.94	15.66	87
5	0.52	8.75	95.76
6	0.25	4.24	100

Table 1: Results of PCA

If we stick to the 67% of variance explained, we get that we should extract the first three components. Therefore, we will end up working with half of the original dimensions and with a 71.74% of the total variance explained. If we based the decision on extracting those dimensions whose eigenvalue is greater than one, we would get the same result, so it seems pretty valid to take just the first three components.

In order to analyze the loadings/correlations, we need to take into consideration the direction of the vectors (variables projected in the same axis are highly correlated - positively or negatively depending on the direction) and the length of the vectors (longer vectors mean higher correlation).

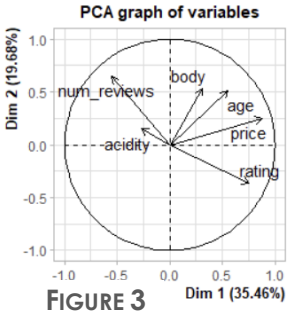

FIGURE 3

On the one hand, we see that the first dimension is highly correlated with price, rating, and age. Acidity is also correlated, although to a lesser extent and negatively. For the second component, we see that body is highly positively correlated, whilst num_reviews also seems to have a strong positive correlation.

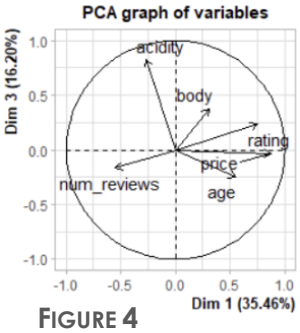

FIGURE 4

The third dimension confirms the fact that acidity wasn't correlated at all with the first component, as it appears to be strongly positively correlated with the third dimension. For the other variables, though, it does not look like there is a clear relation with this dimension.

If we take a look at the contributions from each variable to each dimension, the assumptions we made seem to be correct.

	Dim 1	Dim 2	Dim 3
Rating	26.78	11.12	5.58
Num_reviews	14.25	35.43	2.83
Price	36.58	5.21	0.10
Body	4.62	24.08	14.61
Acidity	3.34	1.90	70.7
Age	14.39	22.23	6.15

Table 2: Contributions of PCA

When applying PCA, it can be useful to label the different dimensions, as they will become the new variables to work with.

The first dimension can be named overall wine quality; it is strongly correlated with price, as well as rating, and less with the age. For instance, it is well known that old wines are more appraised by the wine sommeliers, so it makes sense that exquisite and premium wines tend to be expensive. It is important to remark that it looks like, as the number of reviews is slightly negatively related to the first dimension, that as the number of reviews increases, it has a negative impact on the rating. A hypothesis of this phenomenon could be that cheaper wines, thus "worse wines," are more accessible to the public and therefore have a larger number of ratings.

The second dimension has two main correlations; more than half of the dimension is explained by body, so we could name this dimension like that. However, the number of reviews also seems to exert an important negative influence on the dimension.

The third and last dimension is mainly related to acidity, as it contributes to 63.5% of the dimension. Therefore, we will name the dimension acidity.

Let's check if the analysis is correct by analyzing specific individuals:

Although the plot seems dense, we focus on extreme observations to extract information. Most data clusters around lower ratings, with few observations extending positively along the first dimension, representing top-rated and expensive wines. Old wines are also positioned positively but contribute less than price and rating. Observations are uniformly distributed across the second dimension, indicating a variety of wine bodies and different numbers of votes.

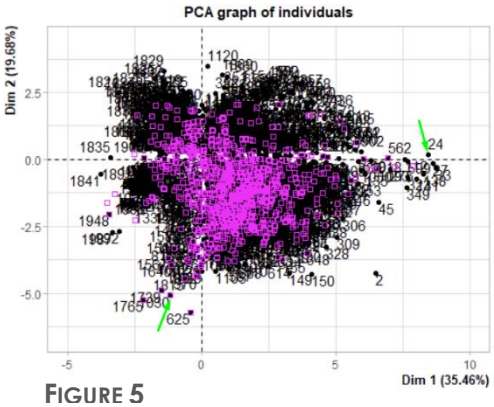

FIGURE 5

For example, observation 24 corresponds to the expensive, highly rated, and old Pedro Ximenez wine, which has few ratings and is appropriately placed. In contrast, observation 1030 is a cheap, young cava from Freixenet. The dataset is sorted by rating in descending order, with wines further right (more expensive and better rated) having lower indices, while those on the left have higher indices.

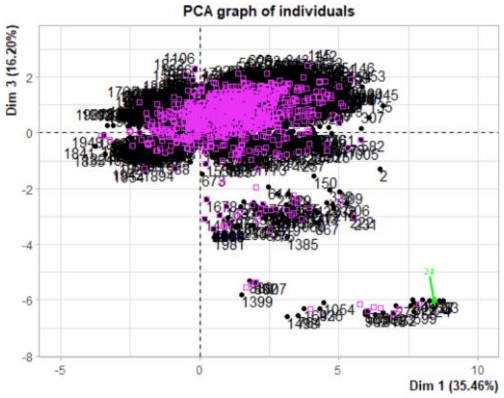

FIGURE 6

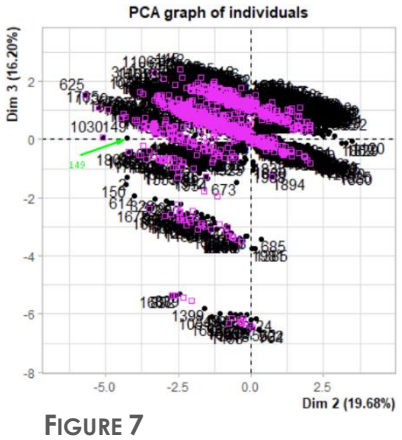

FIGURE 7

What we see in this graphic is easy to understand; dimension 3 is mainly related to acidity. However, almost every wine has an acidity of 3 (mean is 2.947), so the big majority of the samples are forming a cloud. The few observations that are further are those that have lower acidity. One example could be Pedro Ximenez's wine Toro Albala Don PX Convento Seleccion, which has an acidity value of 1. Another reason why it is located on the right is because of its price, rating, and age; it is not a cheap wine, it has a good rating (4.8, the second highest possible), and it is almost a hundred years old, so it makes sense it's in the bottom right.

When looking at individuals in dimensions 2 and 3, we see again that those that have low acidity are far from the other observations, as they are a "rare" observation. When it comes to dimension 2, that was mainly explained positively by body and negatively by the number of observations, we can see different observations: Observation number 149, Recaredo Reserva Particular, a cava that has a low body rating and a number of observations lower than the mean (35 to 404), so it's placed in the left corner. In the right corner, we see the 10th observation with 630 reviews, higher than the average, and the highest rating for body.

MDS

The MDS technique is used to analyze similarity data, understanding patterns and relationships within the dataset through a lower-dimensional representation of the data (in this particular case using only two dimensions). As previously mentioned, for this particular method, it would be beneficial to treat the variable "rating" as categorical to enable a more effective analysis.

So, let's visualize the relations that can be found in the dataset. For this first analysis, only numerical variables are going to be used to calculate the distance matrix.

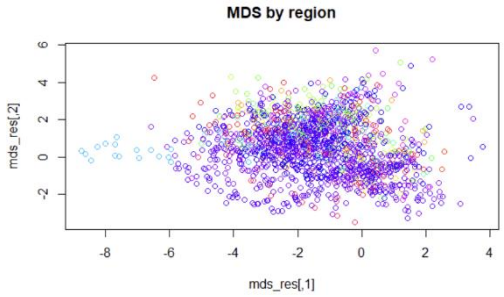

FIGURE 8

From a first sight, no distinct clusters based on the wine's region are evident. The plot displays all the colours of the rainbow mixed. It would be beneficial to have wines clustered by regions, as wines from different regions may possess distinct properties. Nevertheless, an interpretation of this initial MDS plot can still be made.

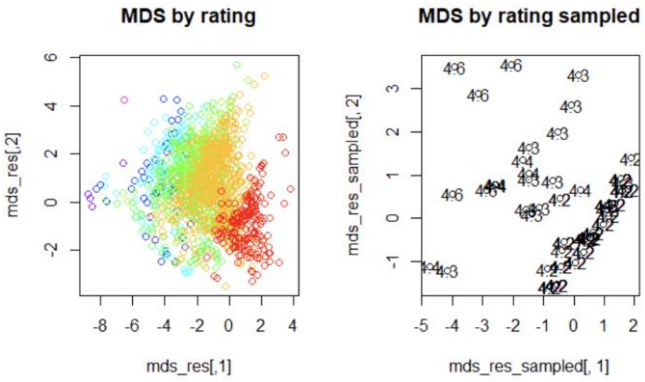

FIGURE 9

By assigning each data entry a different color based on its rating, it becomes apparent that they form clusters, although these clusters may overlap. Labeling only a subset of the MDS results enables a clearer interpretation of the plot, as labeling all entries would result in overlapping labels, obscuring any meaningful interpretation. It is observable that wines of lower quality cluster closer together, while they are positioned far from wines of higher quality. In this particular case, negative values of the first dimension are associated with better quality.

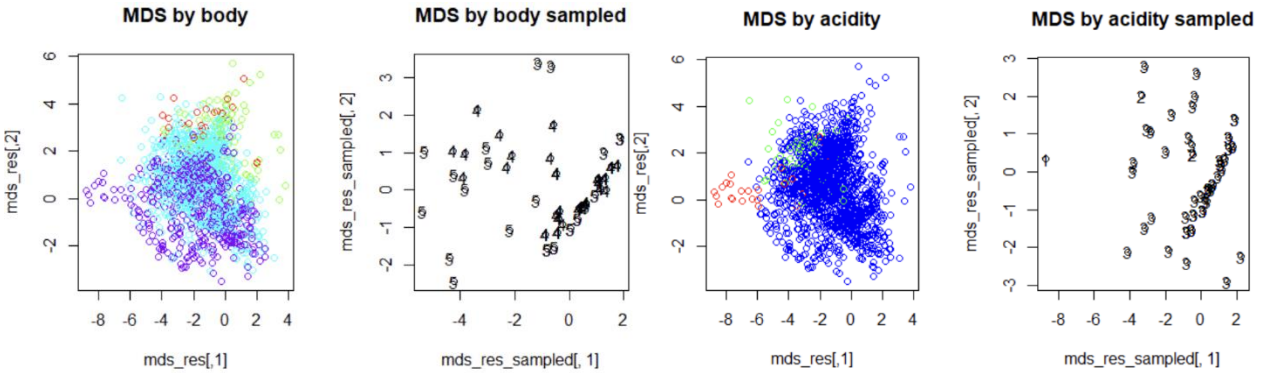

FIGURE 10

FIGURE 11

Similar plots are obtained when colouring by age.

The y-axis represents the clustering of data based on the body of the wine, where higher values on the y-axis correspond to lower values of body. Furthermore, it is evident that different wines cluster depending on their acidity.

In conclusion, following an initial analysis, the application of the MDS technique on a distance matrix derived from numerical variables reveals that wines with similar properties, and thus similar quality, tend to cluster together.

Let's now calculate the distance matrix using as well a categorical variable: the wine type.

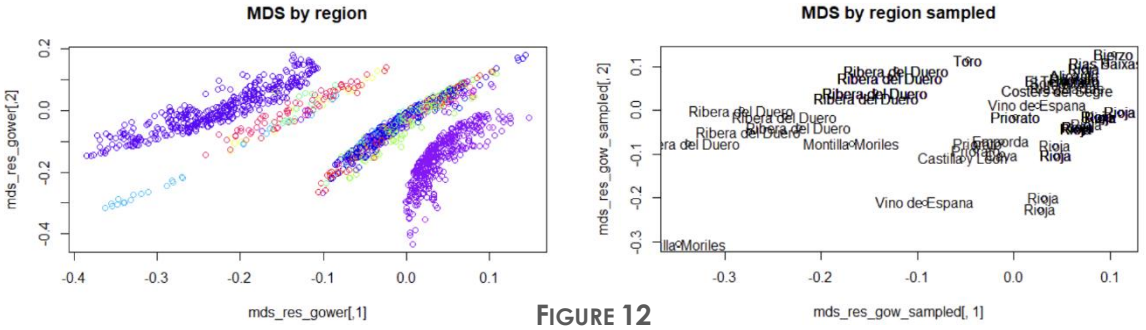

FIGURE 12

Five distinct clusters are clearly identifiable. The cluster on the left represents wines from Ribera del Duero, while the cluster on the right corresponds to wines from La Rioja. The distance between these two clusters in the plot is translated in differences in the wines, specially in their types. Although they use mainly the same type of grape (see wine guide (wineguide)) different winery traditions are used in different regions making wines differ.

The two clusters in the middle encompass various regions, making interpretation challenging. These regions likely share similar weather conditions and grape varieties.

Furthermore, a small cluster (light blue) can be observed at the bottom left, corresponding to wines from Montilla Moriles.

In conclusion, after evaluating the proximity of different wines based as well on their type, it is evident that they tend to form clusters based on their regions. Wines from the same region exhibit greater similarity in type.

CA

In this section we will use Correspondence Analysis to understand relationships between some of our categorical variables. If we try to do so, we encounter some problems because variables have many categories and that makes is hard to interpretation.

The left image is an example of our problem. Here we have tried to apply Correspondence Analysis between type and region, but we have so many regions that we can't visualize the plot correctly.

However, our most important variable may be 'rating' (numerical) and we would like to analyze relationships between this feature and others as well. For this reason, we have decided to make 'rating' a categorical value by creating some tiers.

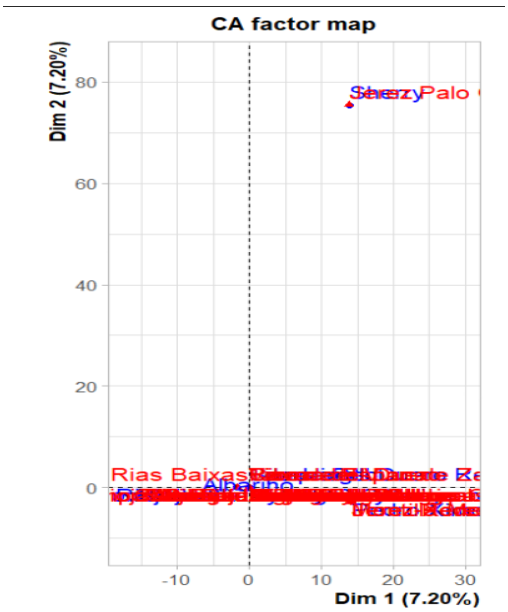

FIGURE 13

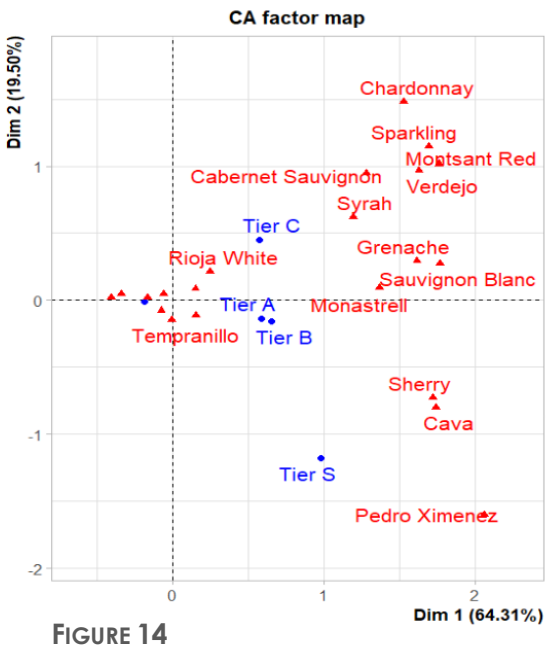

FIGURE 14

Now we have a more visual and interpretable plot, as we can see in the right image. The first that gets our attention is that Dimension 1 explains 64.3% of the variability and Dimension 2, 19.50%, so the first dimension explains three times more than the second.

We can see this in the plot by comparing tiers. In the plot, Tier S and Tier D (the blue point without label) are separated from Tiers A, B and C. For this reason, we interpret that types in A, B and C are very similar and different from the best and worst types of wines. Therefore, we think that Dimension 1 is separating the best rated wines from the worst, while the ones in the middle are merged. This makes sense as the range of the previous numeric 'rating' feature only took values from 4.2 to 4.9.

Looking at types of wines instead of tiers, we can also see this pattern. Pedro Ximenez is separated from the rest as it is the most rated type of wine. Sherry and Cava are very similar, that is the reason why they are close to each other in this dimension. On

the other hand, Rioja White is closer to 0 and far from the ones just mentioned above because his rating is near 4.2, which is a bad rating knowing the range of values. So, in terms of columns (types of wines), we are also separating the best rated from the worst.

Dimension 2 is difficult to interpret because it only explains 20% of the variability. From this dimension we can extract some relations between categories. We can see that Sherry and Cava are very close, as they have similar rating, but far from Pedro Ximenez in comparison with Dimension 1. Chardonnay and Sparkling are closer as well because of having a similar rating and price. These two are far from Sherry and Cava because of his different price. Despite of that, we can't find any patterns here, as we find similar values in types that are far in terms of Dimension 2, and different values in types that are closer.

We have the same problem with tiers. We see that Tier S is far from the rest, which is good, but the others are close to each other, leading to a difficult interpretation. Therefore, we consider that Dimension 2 is not giving much importance to rating as it does Dimension 1.

Clustering

In this section our aim is to group similar observations by means of some variables. In this case, we encounter problems as we have so many observations that we can't represent its clusters.

We have already transformed rating to a categorical variable, but clustering tiers does not give us much information as there are few of them. Alternatively, we can cluster types of wines, as there only are 21 and each one represents very well the wines that includes.

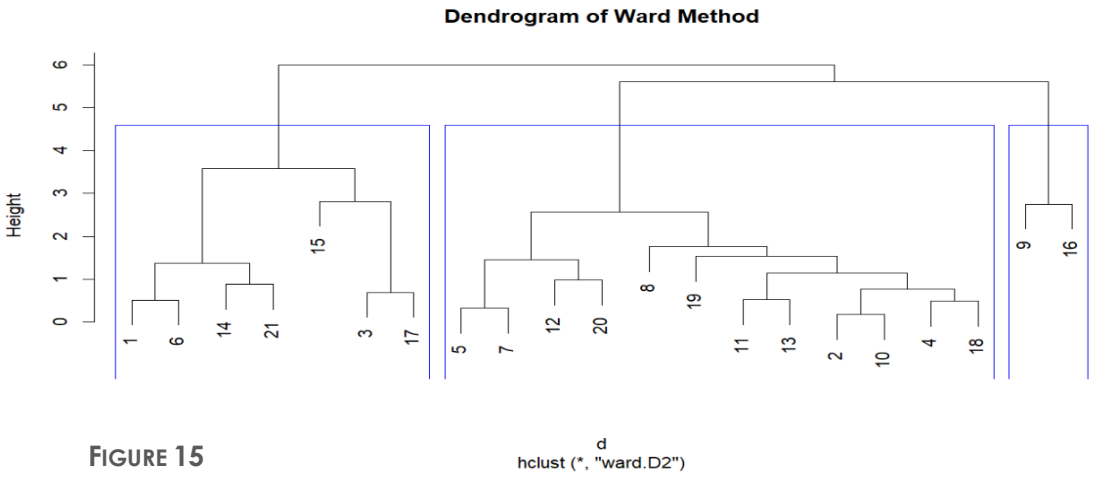

FIGURE 15

Ward is the method which gives us more homogeneaus and interpretable groups. The main difference between using 2 and 3 clusters is separating types of number 9 and number 16 in other group. Our first thought about that is thinking they are outliers, but in this case, we won't treat them in that way as we extracted types by averaging

numeric features of the wines. For this reason, we don't care having two types of wines in one cluster, we interpret them as two exceptional types of wines and separated from the rest.

If we perform a more in-depth analysis, we see that types 9 and 16 are 'Pedro Ximenez' and 'Sherry', the two most expensive and well-rated types of wines, so we want to keep them in the study for a better interpretation.

However, we'll represent the elbow plot to see which number of clusters is the optimum in this case.

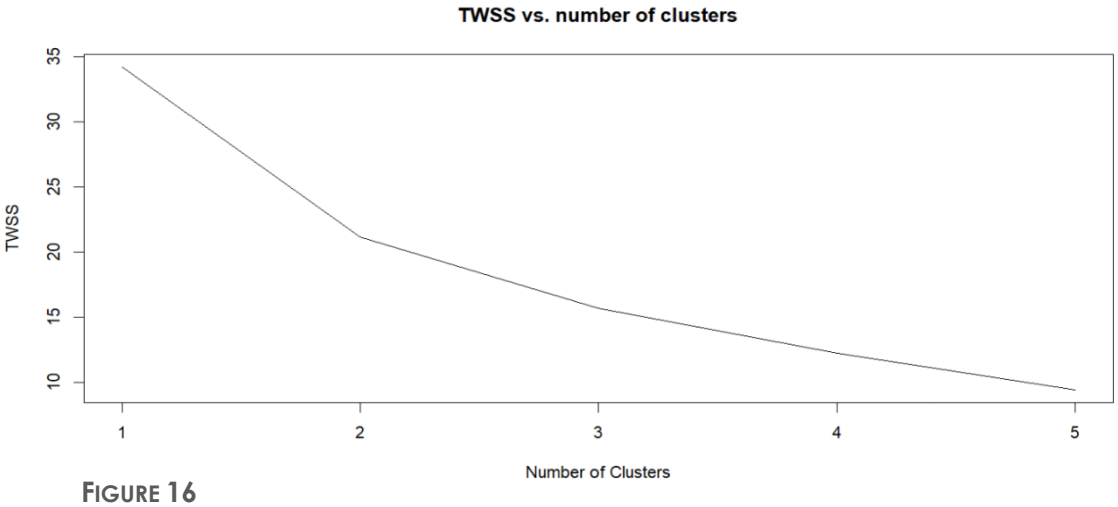

FIGURE 16

As we see in the elbow plot, there are two "elbows" in 2 and 3 number of clusters. Despite of observing a more drastic change in slope using 2 clusters, we prefer using 3 to separate the two best rated wines from the rest, as we commented previously.

Group	Rating	Num_rev	Price	Body	Acidity	Age
1	4.50	3.81	4.99	4.5	2	3.47
2	4.3	5.42	3.72	4.33	2.91	1.81
3	4.32	4.87	3.30	2.71	3	1.73

Table 3: Center mean of each cluster

From this table we can see the different characteristics of each cluster. Group 1 has the best-rated types of wines, as his average rating is far from the other groups. However, let's analyze each group precisely.

On the one hand, types in group 1 have the oldest wines, so here we can see a relation between rating or price with age of wines. In terms of body, wines in group 1 are the best as well, but they seem to be less acid than the rest. Another observation is that the most-rated wines have the least number of reviews in average, maybe because they are the most expensive ones. Therefore, from group 1 we interpret positive correlation between rating and price/body/age, and we see negative correlation between rating and acidity/number of reviews.

On the other hand, groups 2 and 3 have a more difficult interpretation. In group 2, wines are better rated than in group 3, but cheaper and younger as well. Moreover, group 3 stands out in terms of body whereas group 2 is more acid. For this reason, we think that cluster 2 and 3 are separating types of wines in terms of body and acidity rather than rating, as they are very close in this aspect.

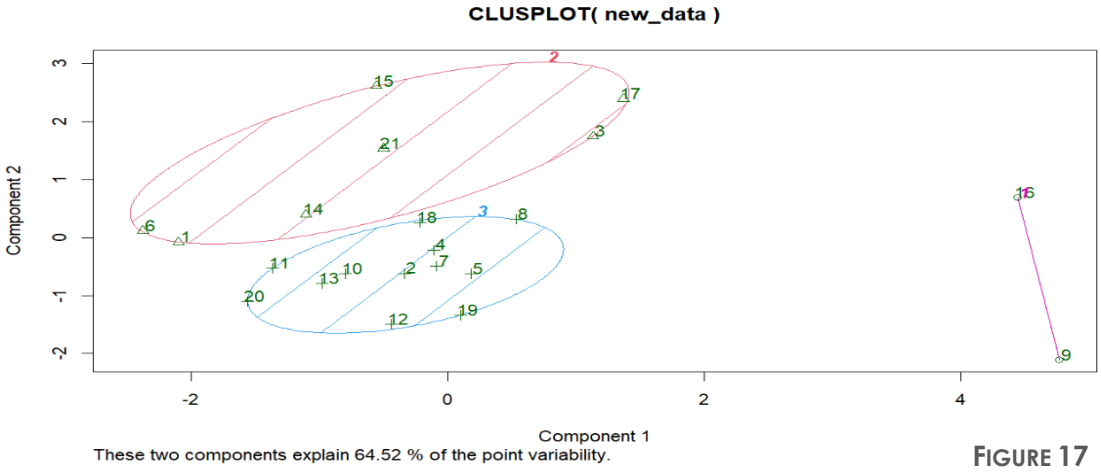

FIGURE 17

Plotting clusters gives a visual interpretation of all the analysis. The first observation is that clara method doesn't help us on the visualization task because all cluster are overlapped, and we want them separated from each other. Instead, K-means (the one used in the image) separates the three clusters very well. From plot we can validate our hypothesis that 9 and 16 (from cluster 1) correspond to best and most rare type of wines whereas cluster 2 and 3 have types that are closer to each other but differ in some characteristics different from the rating.

Discriminant Analysis

We will use Discriminant analysis to predict the rating with all the other data. To begin with, we need to make some assumptions; Normality and Homogeneity of Variance, but I was already checked in the preprocessing stage, so we do not have to worry.

We have different type of predictors, such as LDA, QDA and Naïve Bayes Classification. We have a problem and is caused by the fact that we only have one observation. This makes it impossible for QDA to operate and makes it harder for other predictors. For this reason, we will change its rating to 4.8.

Even though we removed the class with just one observation it will be difficult to make good predictions, as classes are highly unbalanced:

Rating	4.2	4.3	4.4	4.5	4.6	4.7	4.8
Freq	4544	522	392	254	141	51	10

Table 4: Frequencies of each rating

Therefore, LDA gives great results for 4.2 class but poor for the others. However, this result is not improved by QDA nor Naïve Bayes. Given this result, we assume that the best classifier for our data is LDA, that gives a total accuracy of 0.8167061.

However, none of the predictors seem useful as they lack accuracy on lower populated classes.

Appendix

Appendix A

The RMD file included in the folder – `AD-PROJECT.Rmd` – contains all the code and results analyzed in the report.

Moreover, after each output, there is the same explanation made in this report or one more elaborated. What is more, all the tables written on this report are approximations of the results obtained with the code but with a more visual format.

Finally, we would like to inform about the running time of the code. As the dataset is quite big (7500 observations) calculating the distance matrix using either Euclidian or Gower metrics, takes a long time. Patience is required.

Bibliography

(s.f.). Obtained from https://www.kaggle.com/datasets/fedesoriano/spanish-wine-quality-dataset

(s.f.). Obtained from https://ohmygoodguide.com/es/guia-vinos-espana/